PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-119791
(43)Date of publication of application : 30.04.1999

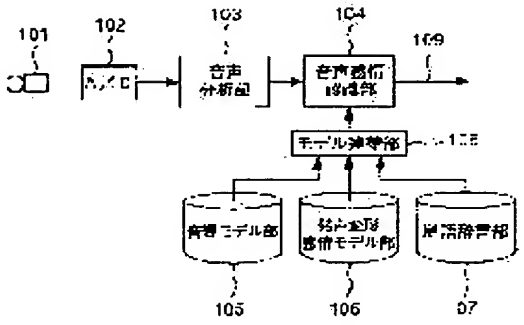
(51)Int.CI. G10L 3/00
G10L 3/00
G10L 3/00
// A63F 9/22
G06F 17/28

(21)Application number : 09-286372
(22)Date of filing : 20.10.1997

(71)Applicant : HITACHI LTD
HITACHI ULSI SYSTEMS CO LTD
(72)Inventor : WAKIZAKA SHINJI
KONDO KAZUO
OBUCHI YASUNARI
TOUSHITA TETSUJI
ISHIKAWA YASUYO

(54) SYSTEM AND METHOD FOR VOICE FEELING RECOGNITION

(57)Abstract:
PROBLEM TO BE SOLVED: To recognize the level of a speaker's feeling by a voice recognition system.
SOLUTION: The system and method are equipped with a dictionary part 107 where object words of voice recognition are gathered, a voice analysis part 103 which performs a voice analyzing process, a sound model part 105 which has patterns of voice in phoneme units, a voicing deformation feeling model part 106 which represents the deformation of a vocal sound spectrum by a feeling, and a voice recognition part 104 which performs a voice recognizing process by coupling the sound model part 105, voicing deformation feeling model part 106, and dictionary part 107, and outputs an object word of voice recognition as a voice recognition result and also outputs a feeling level representing the degree of the speaker's feeling that the voice has. Other voice analysis parts output feeling levels from the features of the power of the voice.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－119791

(43)公開日　平成11年(1999)４月30日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ１０Ｌ　3/00	５３１	Ｇ１０Ｌ　3/00	５３１Ｎ
			Ｒ
	５３５		５３５
// Ａ６３Ｆ　9/22		Ａ６３Ｆ　9/22	Ｆ
Ｇ０６Ｆ　17/28		Ｇ０６Ｆ　15/38	Ｖ

審査請求　未請求　請求項の数８　ＯＬ　（全 13 頁）

(21)出願番号　特願平9－286372

(22)出願日　平成９年(1997)10月20日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(71)出願人　000233169
株式会社日立超エル・エス・アイ・システムズ
東京都小平市上水本町５丁目22番１号

(72)発明者　脇坂　新路
東京都小平市上水本町五丁目20番１号株式会社日立製作所半導体事業部内

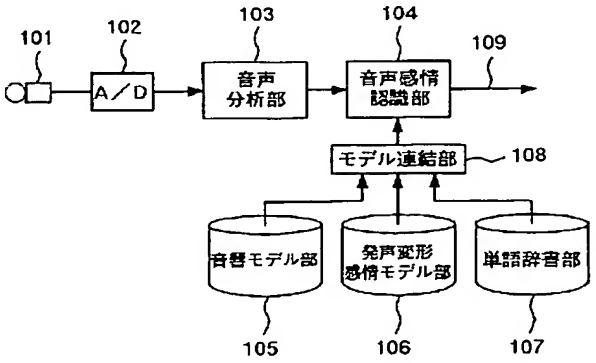
(74)代理人　弁理士　高橋　明夫　（外１名）

最終頁に続く

(54)【発明の名称】　音声感情認識システムおよび方法

(57)【要約】
【課題】音声認識システムにおいて話者の感情のレベルを認識する。
【解決手段】音声認識の対象となる単語を集めた辞書部と、音声分析処理を行う音声分析部と、音声のパターンを音素単位でもつ音響モデル部と、感情による音韻スペクトルの変形を表す発声変形感情モデル部と、音声分析結果に対して、音響モデル部と発声変形感情モデル部と辞書部とを連結して音声認識処理を行う音声認識部とを備え、音声の特徴から、音声認識の対象となる単語を音声認識結果として出力すると共に、音声がもっている話者の感情の度合を示す感情レベルを出力する。他の音声分析部は音声のパワーの特徴から感情レベルを出力する。

図　1



【特許請求の範囲】
【請求項１】音声認識の対象となる単語や文章を集めて辞書として定義し、音声認識結果として、それらの単語や文章を辞書部からピックアップして、文字列表示や音声合成を用いて出力する音声認識システムにおいて、
取り込んだ音声に対して音声分析処理を行う音声分析部と、音声のパターンを音素単位でもつ音響モデル部と、感情による音韻スペクトルの変形を表す発声変形感情モデル部と、音声分析結果に対して、音響モデル部と発声変形感情モデル部と辞書部とを連結して、音声認識処理を行う音声認識部とを備え、音声の特徴から、音声認識の対象となる単語や文章を音声認識結果として出力するとともに、音声がもっている話者の感情の度合を示すレベルを出力することを特徴とする音声感情認識システム。
【請求項２】請求項１記載の音声感情認識システムにおいて、
音声が持っている話者の感情の度合を示すレベルは、数字０～N（Nは整数）であることを特徴とする音声感情認識システム。
【請求項３】請求項１記載の音声感情認識システムにおいて、
前記辞書部は、音声認識の対象となる単語や文章を集めた辞書と、それらの単語や文章に対して、音声が持っている感情のレベルを表現する修飾語を集めた辞書とを備え、音声認識結果として、それらの単語や文章をピックアップするとともに、感情のレベルを表現する修飾語をピックアップして、単語や文章に修飾語を付加して、文字や音声合成を用いて出力することを特徴とする音声感情認識システム。
【請求項４】音声認識の対象となる単語や文章を集めて辞書として定義し、音声認識結果として、それらの単語や文章をピックアップして、文字列表示や音声合成を用いて出力する音声認識システムにおいて、
取り込んだ音声に対して音声分析処理を行う音声分析部と、音声のパターンを音素単位でもつ音響モデル部と、音声分析結果に対して音響モデル部と辞書部とを連結して音声認識処理を行う音声認識部とを備え、取り込んだ音声に対して、音声分析処理を行う音声分析部は、感情の度合が現われる音の強弱を示すパワーの特徴から、感情の度合を示すレベルを出力することを特徴とする音声感情認識システム。
【請求項５】音声認識の対象となる単語や文章を集めた辞書と、取り込んだ音声に対して音声分析処理を行う音声分析部と、音声のパターンを音素単位でもつ音響モデル部と、感情による音韻スペクトルの変形を表す発声変形感情モデル部と、音声分析結果に対して、音響モデル部と発声変形感情モデル部と辞書部とを連結して音声認識処理を行う音声認識部とを備え、音声の特徴から音声認識の対象となる単語や文章を音声認識結果として出力するとともに、発声変形感情モデル部からのデータを用いて音声がもっている話者の感情の度合を出力することを特徴とする音声感情認識方法。
【請求項６】音声認識の対象となる単語や文章を集めて辞書と、取り込んだ音声に対して音声分析処理を行うと共に音のパワーを分析して感情のレベルを出力することが出来る音声分析部と、音声のパターンを音素単位でもつ音響モデル部と、音声分析結果に対して音響モデル部と辞書を連結して音声認識処理を行う音声認識部とを備え、音声分析部は取り込んだ音声に対して感情の度合が現われる音の強弱を示すパワーの特徴から、感情の度合を認識し、この出力感情のレベルを出力することを特徴とする音声感情認識方法。
【請求項７】請求項５又は６記載の音声感情認識方法において、
音声が持っている話者の感情の度合を示すレベルは、数字０～N（Nは整数）であることを特徴とする音声感情認識方法。
【請求項８】請求項５又は６記載の音声感情認識方法に於いて、
前記辞書部は、音声認識の対象となる単語や文章を集めた辞書と、それらの単語や文章に対して、音声が持っている感情のレベルを表現する修飾語を集めた辞書とを備え、音声認識結果として、それらの単語や文章をピックアップするとともに、感情のレベルを表現する修飾語をピックアップして、単語や文章に修飾語を付加して、文字や音声合成を用いて出力することを特徴とする音声感情認識システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、音声認識システムおよび方法に係わり、カーナビゲーションシステム、車載用PC、PDA（パーソナル・デジタル・アシスタント）、ハンドヘルドPCに代表される小型情報機器、携帯型音声翻訳機、ゲーム、家電機器に用いる音声認識システムであって、特に、音声認識の対象となる単語や文章の認識とともに、感情を表わす単語や文章においては、感情の度合まで認識する音声感情認識システムおよび方法に関する。
【０００２】
【従来の技術】近年、音声認識技術を用いた小型情報システムが普久しつつある。カーナビゲーションシステムをはじめとして、PDAに代表される小型情報機器、携帯型翻訳機等である。
【０００３】このような音声認識システムの例として、特願平５－３５７７６号公報の「言語自動選択機能付翻訳装置」には、マイクから入力した操作者の音声を認識して、翻訳し、翻訳した言語の音声を出力するようにした携帯用の翻訳装置に関する技術が開示されている。
【０００４】以下、図７を用いてこのような従来技術に

係わる音声翻訳装置の概要について説明する。

【０００５】図７は従来技術に係わる音声翻訳装置の構成を示すブロック図である。制御部７０１は、マイクロプロセッサ等からなり、装置の各部を制御する。音声区間切出し部７０２は、マイク７０９から入力された音声をデジタル信号に変換して切り出し、音声認識部７０３に送る。音声認識部７０３は、キーボード又はスイッチ等による操作信号７１１を受けた制御部７０１の指示により、マイク７０９、音声区間切出し部７０２を経て、切り出された音声を分析する。そしてその結果を、音声認識辞書部７０７に格納された標準音声パターンと比較することにより、音声認識をおこなう。音声合成部７０５は、音声認識部７０３により認識された音声に対応した翻訳語を、翻訳語データ用メモリカード７０６から読み込み、音声信号に変換してスピーカアンプ７１０、スピーカ７０８を経て出力する。

【０００６】表示部７０４は、翻訳装置の使用者への指示や翻訳語の文字による表示等をおこなう。翻訳語データ用メモリカード７０６は、ＲＯＭカード等からなり、翻訳語を音声合成して出力する場合には、音声データを格納している。また、この翻訳語データ用メモリカード７０６から、翻訳語に対応したキャラクターコードを読み込み、表示部７０４に表示する。そして、この翻訳語データ用メモリカード７０６を他の言語のものと交換することにより、複数の言語に翻訳することが可能となる。音声認識辞書部７０７は、ＲＡＭ等からなり、操作者の発生に応じた標準音声パターンを格納している。この標準音声パターンは、操作者があらかじめ格納しておく。

【０００７】

【発明が解決しようとする課題】このような音声認識、音声合成技術の分野は、半導体技術の向上を背景として、システムがより人間的なユーザインタフェースを提供すべきであるという要望から、その発展が期待されている。上記従来の音声認識技術を用いた小型情報システムにおいても、カーナビゲーションシステムをはじめとして、ＰＤＡに代表される携帯型情報機器、携帯型翻訳機、さらに、音声インタフェースを持った情報家電として、今後ますます普及してくることが予想される。

【０００８】しかしながら、音声認識は、処理すべき情報量が膨大なものになるため、従来の技術では、認識率や認識応答時間の性能を低下させないためには、認識する語数に制約を設ける必要がある。そのためには、あらかじめ登録しておいた単語、文に対して、その文字列が持つ統計的な話者の音声の特徴と、実際に話者が発声した音声の特徴とを比較し、確率的に一番近い値を認識結果としている。

【０００９】今後、音声認識における技術革新や、それを実現するソフトウエア、ハードウエアの性能向上により、認識率や認識応答時間の性能は向上することが考えられる。そこで、さらに、人間的なユーザインタフェースを提供するためには、単に、従来の音声認識技術において、あらかじめ登録した単語、文の文字列を認識するだけでなく、話者の感情や意図を認識できれば、たとえ制限された認識語数においても、使い勝手の向上が期待できる。しかしながら、従来の音声認識システムでは、あらかじめ登録した単語や文の文字列のみを音声で照合して、入力した音声に最も近い文字列を音声認識結果として出力する音声認識システムであり、音声を発声した話者の感情や意図までは認識できない。

【００１０】本発明は、システムが少しでも人間的なユーザインタフェースを持てるように、上記問題点を解決するためになされたものである。

【００１１】本発明の目的は、小型情報システムに用いられる音声認識システムにおいて、入力された音声に対して、辞書に登録された単語や文の文字列を認識するとともに、入力された音声が持っている話者の感情や意図を認識することができる音声感情認識システム及び方法を提供することにある。

【００１２】また、本発明の他の目的は、小型情報システムに用いられる音声認識システムにおいて、入力された音声が持っている話者の感情や意図を感情の度合いを表現する数字や修飾語に変換して、人間とシステムにおける良好な音声インタフェースを実現することである。

【００１３】

【課題を解決するための手段】上記目的を達成するために、本発明の音声感情認識システム及び方法に於いては音声認識の対象となる単語や文章を集めて辞書として定義し、音声認識結果として、それらの単語や文章を辞書部からピックアップして、文字列表示や音声合成を用いて出力する音声認識システムにおいて、取り込んだ音声に対して音声分析処理を行う音声分析部と、音声のパターンを音素単位でもつ音響モデル部と、感情による音韻スペクトルの変形を表す発声変形感情モデル部と、音声分析結果に対して音響モデル部と発声変形感情モデル部と辞書部とを連結して音声認識処理を行う音声認識部とを備え、音声の特徴から音声認識の対象となる単語や文章を音声認識結果として出力するとともに、音声がもっている話者の感情の度合を出力するようにしたものである。

【００１４】より詳しい１実施例に於いては、音声が持っている話者の感情の度合を示すレベルは、数字０～Ｎ（Ｎは整数）であるようにしたものである。

【００１５】また、本発明の音声感情認識システム及び方法に於いては、音声が持っている話者の感情の度合は、音声認識の対象となる単語や文章を集めた辞書とそれらの単語や文章に対して感情のレベルを表現する修飾語を集めた辞書と有する辞書部を備え、音声認識結果として、それらの単語や文章をピックアップするとともに、感情のレベルを表現する修飾語をピックアップし

て、単語や文章に修飾語を付加して、文字や音声合成を用いて出力するようにしたものである。

【００１６】さらに詳しい１実施例に於いては、音声認識の対象となる単語や文章を集めて辞書として定義し、音声認識結果として、それらの単語や文章をピックアップして、文字列表示や音声合成を用いて出力する音声認識システムにおいて、取り込んだ音声に対して音声分析処理を行う音声分析部と、音声のパターンを音素単位でもつ音響モデル部と、音声分析結果に対して音響モデル部と辞書部とを連結して音声認識処理を行う音声認識部とを備え、取り込んだ音声に対して音声分析処理を行う音声分析部は、感情の度合が現われる音の強弱を示すパワーの特徴から感情の度合を出力することが出来る。

【００１７】

【発明の実施の形態】以下、本発明に係る各実施形態を図 １から図６を用いて説明する。

【００１８】図１は本発明に係る音声および感情認識システムの各機能とその処理の流れを示すブロック図である。

【００１９】音声および感情認識をおこなうために、図１に示されるマイク１０１から音声が取り込まれる。取り込まれた音声であるアナログ信号は、アナログ信号をデジタル信号に変換するＡ／Ｄ変換器１０２によって、任意に決められたサンプリング周期により、アナログデータからデジタルデータに変換される。変換された音声のデジタルデータは、音声分析部１０３によって、雑音処理や音声分析や話者適応などの前処理がなされ、音声感情認識部１０４により音声および感情認識がなされる。ここで、音声および感情認識とは、２つの処理を実行する。

【００２０】第１の処理は、音声信号を解析して、それを短い時間ごとの音素として分析して、そのパターンを解析し、該当する単語や文章を辞書から選択することである。

【００２１】第２の処理は、音声信号を解析して、それを短い時間（５～２０ｍｓ）ごとの音素として分析して、そのパターンを解析し、話者が発声した音声の感情の度合いを示すレベルを単語や文章ごとに選択することである。

【００２２】以上の２つの処理から、音声感情認識システムの出力として、音声認識結果および音声の感情レベル１０９を生成する。

【００２３】音声感情認識部１０４は、音声分析部１０３で分析された入力音声の音声分析結果に対して、音響モデル部１０５、発声変形感情モデル部１０６、単語辞書部１０７をモデル連結部１０８によって連結された音素単位で照合して、単語辞書部１０７に登録した単語の中で、一番近い単語をピックアップする。さらに、ピックアップされた単語の入力音声を持っている感情の度合いを示すレベルを選択する。なを、図１に示す実施例に於いては、電源を投入すると、モデル連結部１０８で連結された音素単位の単語及び感情の度合いを示すレベルは音声感情認識部１０４に記憶され、音声分析部１０３からの音声分析結果と直ちに照合出来るようになっている。

【００２４】音響モデル部１０５は、音声認識に用いられるモデルであり、具体的には、単語辞書部１０７に用いられている文字と音素との対応であり、音素の特徴が出現する確率の分布、出現した音素の特徴が次のどの特徴が現れる状態に遷移するかの確率の分布を記憶したものである。音素の特徴が出現する確率の分布について説明する。例えば、「あつい」という音声の「あ」に対して音声スペクトラムは人によって変わるため、「あ」と言う音素に対して、横軸に音声スペクトラムをとり、縦軸に音素が出現する確率を取ると、音声スペクトラムに対して「あ」と認識される確率が変わることを言う。次に、出現した音素の特徴が次のどの特徴が現れる状態に遷移するかの確率の分布について説明する。例えば、「あ」という音素は「あつい」のように次に「つ」がくる場合もあるし、「あさい」のように次に「さ」がくることもあるし、「あまい」のように次に「ま」に遷移することもある。「あ」が次にどの音素に遷移するかの確率は各音素によって変わる。つまり、ある音素の特徴が次にどの音素の特徴に変化するかの確率は変わるので、この確率の分布を言う。

【００２５】音響モデル部１０５は、あらかじめ声を登録しなくても、誰が話し手でもその声を認識できるいわゆる「不特定話者対応」が、一般的になってきている。このような音響モデルとしては、例えば、隠れマルコフモデル（ＨＭＭ：Hidden Markov Model）を用いることができる。

【００２６】発声変形感情モデル部１０６は、感情の変化による音韻スペクトルの変形要素に着目して、感情が変化したときの単語辞書部１０７に用いられている文字と音素との対応である。即ち、感情を込めたときに音素の確率の分布が変わるが、その時の音素の特徴が出現する確率の分布、出現した音素の特徴が次のどの特徴が現れる状態に遷移するかの確率の分布を記憶したものである。この、出現した音素の特徴が次のどの特徴が現れる状態に遷移するかの確率の分布は、例えば、「あつい」と言う言葉を感情を込めて「あつい」と言っても変化しないが、「あちー」とか「あちい」に変化した場合に変わる。このような発声変形感情モデル部１０６としては、例えば、隠れマルコフモデル（ＨＭＭ：Hidden Markov Model）を用いることができる。

【００２７】単語辞書部１０７は、言葉、単語（名詞、動詞等）、文章を集めたものである。例えば、カーナビゲーションシステムにおいては、通り名、地名、建造物名、町名、番地、交差点名、個人住宅（個人名）、電話番号等や、必要最小限の会話に必要な言葉の集合体であ

る。ただし、音声認識感情システムでは、特に、単語の中でも、感情を表現する単語、あるいは、感情が現われる単語で構成された単語の集合体である。より具体的には、話者が発声する「暑い」「寒い」「熱い」「冷たい」「はやく」「おそく」「大きい」「小さい」「赤い」「白い」「高く」「低く」「走れ」「進め」「戻れ」「回れ」「飛べ」等の言葉である。また、名詞等の感情表現でない単語も含まれる。この単語辞書部１０７は、システムの能力に応じて一つの辞書あたり、例えば１０～５０００語の単語で構成する。

【００２８】以上から、音声感情認識とは、音声を解析して、それを短い時間ごとの音素として分析して、そのパターンを解析し、該当する単語や文章を辞書から選択するとともに、話者が発声した音声の感情の度合いを示すレベルを単語や文章ごとに選択することである。

【００２９】なお、図１に示す各処理ブロックは、複数のＬＳＩやメモリで構成されたシステムであっても、半導体素子上に構成された一つないし複数のシステムオンチップであってもよい。また、各処理は、専用ＬＳＩや専用ＩＣで処理するハードウエアであっても、ＤＳＰやＲＩＳＣマイコン等のソフトウエアで実現したミドルウエアであってもよい。

【００３０】図２は、隠れマルコフモデル（HMM:Hidden Markov Model）による日本語音素のモデル化の例である。

【００３１】２０１、２０２、２０３は音素分布の状態を表わしている。話者が発声した音声は、「あつい」であり、発音記号の１例で表わす「ａｔｓｕｉ」である。説明を簡単にするために、図２（ａ）において、「ａ」が２０１の状態に対応し、「ｔｓｕ」が２０２の状態に対応し、「ｉ」が２０３の状態に対応している。実際の音声認識では、状態をさらに細分化して表わしている。音声は、非定常信号であり、あるときは「ａ」のスペクトル、あるときは「ｔｓｕ」のスペクトルという具合に、スペクトルの性質が時々刻々と変化することによって言語情報を伝える。この非定常な信号は、性質の異なる定常信号の音素片の連続とみることができる。この性質の異なる定常信号の音素片の一つ一つが、２０１～２０３に示したＨＭＭ状態遷移ネットワークの状態に対応している。この状態、すなわち、非定常信号源からの出力として音声のスペクトルが観測される。観測値は、短時間フレーム毎の音声信号のＬＰＣ分析結果であっても、ベクトル量子化された符号であってもよい。よって、ＨＭＭとは、状態２０１と状態２０２の間の状態遷移確率２０７と、状態２０１から出力される音声のスペクトルが出力される確率２０１である。確率２０１とは、「ａ」の音素分布の内どの確率値が出力されるかと言う事を示す。即ち、２０１～２０３は、音素の分布であり、各状態から出力される音声のスペクトルが出力される確率を示したものである。２０４～２０９は、各状態が次にどの状態に遷移するかの確率を示したものである。この内、２０４～２０６は、例えば、ある音を長く発音したとすると、これを音声分析した場合、ある時間間隔の中では、また同じ音に戻ることを示している。

【００３２】つぎに、音声認識に用いるＨＭＭの一例を説明する。

【００３３】図２（ａ）に示す曲線２１１は、状態２０１から出力される音声のスペクトルが出力される確率を連続分布で表現したものである。ここで、音声のスペクトルは、音声の特徴パラメータをｉ次元としたときのｎ番目の特徴パラメータとする。つまり、音声の特徴を表わす表現方法としては何種類もあるが、仮にこの表現方法がｉ個あったとすると、そのｎ番目の表現方法の特徴パラメータを意味する。横軸は、状態２０１から出力される音声のスペクトルであり、縦軸は、その確率値である。この分布は、平均μ＿ａ、分散σ＿ａをもつ連続分布である。同様に、図２（ｂ）に示す曲線２１２は状態２０２から出力される音声のスペクトルが出力される確率を連続分布で表現したものである。横軸は、状態２０２から出力される音声のスペクトルであり、縦軸は、その確率値である。この分布は、平均μ＿ｔｓｕ、分散σ＿ｔｓｕをもつ連続分布である。

【００３４】図２（ｃ）に示す曲線２１３は、状態２０３から出力される音声のスペクトルが出力される確率を連続分布で表現したものである。横軸は、状態２０３から出力される音声のスペクトルであり、縦軸は、その確率値である。この分布は、平均μ＿ｉ、分散σ＿ｉをもつ連続分布である。

【００３５】ここで、認識対象単語として登録された「ａｔｓｕｉ」の単語辞書部１０７に話者が「あつい」と音声を入力する。「あ」の音声に対して、音声分析が行われ、音声の特徴が出力される。例えば、音声の特徴パラメータをｉ次元としたときのｎ番目の特徴パラメータを使用するものとすると、「あ」の特徴ｆ＿ｎ１が出力される。このとき、単語辞書「ａ」において、特徴ｆ＿ｎ１の出現する確率が連続分布曲線２１１から計算され、確率値ｐ＿ｎ１が出力される。同様にして、「つ」「い」の音声に対して、音声分析が行われ、音声の特徴が出力される。それぞれ、単語辞書「ｔｓｕ」において、特徴ｆ＿ｎ２の出現する確率が連続分布曲線２１２から計算され、確率値ｐ＿ｎ２が出力される。また、単語辞書「ｉ」において、特徴ｆ＿ｎ３の出現する確率が連続分布曲線２１３から計算され、確率値ｐ＿ｎ３が出力される。さらに、音素分布状態２０１から音素分布状態２０２関の状態遷移確率においても同様の処理がおこなわれ、状態遷移先を状態２０７に決定している。最終的に、登録された単語辞書「ａｔｓｕｉ」に対して、音声入力された「あつい」の出現する確率値はＰ＿ａｔｓｕｉ＝ｐ＿ｎ１＋ｐ＿ｎ２＋ｐ＿ｎ３となる。この一連の処理を、登録された単語辞書全てにおいて計算し、確

率値の一番高かったものが、認識結果となる。以上が音声認識の一連の処理である。

【００３６】さらに、図２（ｅ）から図２（ｇ）を用いて、音声感情認識における発声変形感情モデルを用いたHMMの一例を説明する。

【００３７】発声変形感情モデル部１０６は、感情の変化による音韻スペクトルの変形要素に着目して、感情が変化したときの単語辞書部１０７に格納されている単語の文字と音素との対応であり、音素の特徴が出現する確率の分布、出現した音素の特徴が次のどの特徴が現れる状態に遷移するかの確率の分布を記憶したものである。

【００３８】曲線２１１は前に説明したように、状態２０１から出力される音声のスペクトルが出力される確率を連続分布で表現したものである。ここで、音声のスペクトルは、音声の特徴パラメータをｉ次元としたときのｎ番目の特徴パラメータとする。横軸は、状態２０１から出力される音声のスペクトルであり、縦軸は、その確率値である。この分布は、平均μ＿ａ、分散σ＿ａをもつ連続分布である。このとき、音声の特徴パラメータをｉ次元としたときのｎ番目の特徴パラメータにおいて、感情の変化による音韻スペクトルの変形が顕著に現われたとする。そこで、話者が通常の感情で発声したときの音声スペクトルの連続分布曲線を２１１とし、話者が感情をこめて発声したとき、すなわち、感情の変化により変形した時の音声スペクトルの連続分布曲線を２１４とする。よって、従来の音声認識に用いられてきた音響モデルのHMMに加えて、感情の変化により音韻スペクトルの変形が現われる特徴パラメータだけで構成した確率分布を音声感情認識モデルのHMMとして用意する。曲線２１４は、音声感情認識モデルにおいて、状態２０１から出力される感情の変化による音声のスペクトルが出力される確率を連続分布で表現したものである。横軸は、状態２０１から出力される音声のスペクトルであり、縦軸は、その確率値である。この分布は、平均μ＿ａ＿ｓ、分散σ＿ａ＿ｓをもつ連続分布である。ここで、認識対象単語として登録された「ａｔｓｕｉ」の単語辞書において、実際に、話者が「あつい」と感情をこめて音声を入力する。「あ」の音声に対して、音声分析が行われ、音声の特徴が出力される。例えば、音声の特徴パラメータをｉ次元としたときに、ｎ番目の音声の特徴パラメータを採用したとすると、「あ」の特徴ｆ＿ｎ１＿ｅが出力される。このとき、単語辞書「ａ」において、特徴ｆ＿ｎ１＿ｅの出現する確率が連続分布曲線２１４から計算され、確率値ｐ＿ｎ１＿ｅが出力される。ここで、この連続分布曲線２１４に関して、確率値ｐ＿ｎ１＿ｅは、話者が通常の発声をしたときの特徴ｆ＿ｎ１での確率値ｐ＿ｎ１より高い値をとる。

【００３９】また、曲線２２２は、状態２０２から出力される音声のスペクトルが出力される確率を連続分布で表現したものである。横軸は、状態２０２から出力される音声のスペクトルであり、縦軸は、その確率値である。この分布は、平均μ＿ｔｓｕ＿ｓ、分散σ＿ｔｓｕ＿ｓをもつ連続分布である。曲線２２３は、状態２０３から出力される音声のスペクトルが出力される確率を連続分布で表現したものである。横軸は、状態２０３から出力される音声のスペクトルであり、縦軸は、その確率値である。この分布は、平均μ＿ｉ＿ｓ、分散σ＿ｉ＿ｓをもつ連続分布である。

【００４０】曲線２１４の場合と同様に、「つ」「い」の音声に対して、音声分析が行われ、音声の特徴が出力される。それぞれ、単語辞書「ｔｓｕ」において、特徴ｆ＿ｎ２＿ｅの出現する確率が連続分布曲線２２２から計算され、確率値ｐ＿ｎ２＿ｅが出力される。また、単語辞書「ｉ」において、特徴ｆ＿ｎ３＿ｅの出現する確率が連続分布曲線２２３から計算され、確率値ｐ＿ｎ３＿ｅが出力される。さらに、状態と状態の間の状態遷移確率においても同様の処理がおこなわれ、状態遷移先を決定している。最終的に、登録された単語辞書「ａｔｓｕｉ」に対して、感情をこめて音声入力された「あつい」の出現する確率値はＰ＿ａｔｓｕｉ＝ｐ＿ｎ１＿ｅ＋ｐ＿ｎ２＿ｅ＋ｐ＿ｎ３＿ｅとなる。この一連の処理を、登録された単語辞書全てにおいて計算し、計算された確率値の範囲によって感情のレベルを出力する。以上が音声の感情レベルを認識する一連の処理である。

【００４１】図３は、本発明に係る他の音声および感情認識システムの各機能とその処理の流れを示すブロック図である。

【００４２】音声および感情認識をおこなうために、図３に於いては、マイク３０１から音声が取り込まれる。取り込まれた音声であるアナログ信号は、アナログ信号をデジタル信号に変換するＡ／Ｄ変換器３０２によって、任意に決められたサンプリング周期により、アナログデータからデジタルデータに変換される。変換された音声のデジタルデータは、音声分析部３０３によって、雑音処理や音声分析や話者適応などの前処理がなされると共に、音声分析部３０３に含まれている音声パワー分析部３０３ａで音声パワーを分析して感情のレベルが出力される。音声分析部３０３の出力は音声感情認識部３０４で処理され、音声および感情認識がなされる。ここで、音声感情認識部３０４で行われる音声感情認識とは、２つの処理を実行する。

【００４３】第１の処理は、音声信号を解析して、それを短い時間ごとの音素として分析して、そのパターンを解析し、該当する単語や文章を辞書から選択することである。

【００４４】第２の処理は、音声信号を解析して、それを短い時間（５～２０ｍｓ）ごとの音素として分析して、そのパターンを解析し、話者が発声した音声の感情の度合いを示すレベルを単語や文章ごとに選択することである。

【００４５】以上の２つの処理から、音声感情認識システムの出力として、音声認識結果および音声の感情レベル３０９を生成する。

【００４６】音声感情認識部３０４は、音声分析部３０３で分析された入力音声の音声分析結果に対して、音響モデル３０５、単語辞書３０７をモデル連結部３０８によって連結された音素単位で照合して、単語辞書部３０７に登録した単語辞書３０７の中で、一番近い単語をピックアップする。さらに、ピックアップされた単語の入力音声が持っている感情の度合いを示すレベルを選択する。

【００４７】音響モデル部３０５は、音声認識に用いられるモデルであり、具体的には、単語辞書部３０７に格納されている文字と音素との対応であり、音素の特徴が出現する確率の分布、出現した音素の特徴が次のどの特徴が現れる状態に遷移するかの確率の分布を記憶したものである。音響モデル部３０５は、あらかじめ声を登録しなくても、誰が話し手でもその声を認識できるいわゆる「不特定話者対応」が、一般的になってきている。このような音響モデルとしては、例えば、隠れマルコフモデル（ＨＭＭ：Hidden Markov Model）を用いることができる。

【００４８】単語辞書部３０７は、言葉、単語（名詞、動詞等）、文章を集めたものである。例えば、カーナビゲーションシステムにおいては、通り名、地名、建造物名、町名、番地、交差点名、個人住宅（個人名）、電話番号等や、必要最小限の会話に必要な言葉の集合体である。ただし、音声認識感情システムでは、特に、単語の中でも、感情を表現する単語、あるいは、感情が現われる単語で構成された単語の集合体である。より具体的には、話者が発声する「暑い」「寒い」「熱い」「冷たい」「はやく」「おそく」「大きい」「小さい」「赤い」「白い」「高く」「低く」「走れ」「進め」「戻れ」「回れ」「飛べ」等の言葉である。また、名詞等の感情表現でない単語も含まれる。この単語辞書部３０７に格納される単語数は、システムの能力に応じて決められるが、一つの辞書あたり、例えば、１０～５０００語である。

【００４９】以上から、音声感情認識システム又は音声感情認識方法とは、音声信号を解析して、それを短い時間ごとの音素として分析して、そのパターンを解析し、該当する単語や文章を辞書から選択するとともに、話者が発声した音声の感情の度合いを示すレベルを単語や文章ごとに選択することである。

【００５０】なお、図３に示す各処理ブロックは、複数のＬＳＩやメモリで構成されたシステムであっても、半導体素子上に構成された一つないし複数のシステムオンチップであってもよい。また、各処理は、専用ＬＳＩや専用ＩＣで処理するハードウエアであっても、ＤＳＰやＲＩＳＣマイコン等のソフトウエアで実現したミドルウエアであってもよい。

【００５１】図４（ａ）は、図３で説明した音声感情認識システムにおいて、話者が発声した音声「あつい」の音声入力波形を示すもので、横軸は時間を、縦軸は音声レベルを示す。また、図４（ｂ）は「あつい」の音声のパワーを示したものであり、横軸に時間を、縦軸に音声のパワーを示している。

【００５２】音声入力波形４０１は、話者が平常の音声で「あつい」と発声したときの音声波形である。音声信号は、時々刻々と変化する非定常な信号である。この音声信号を２０ｍｓの短時間で切り出して見ると、定常信号と同様なスペクトル音声分析ができる。切り出された音声信号のサンプル値から、例えば、音声分析で広く用いられているＬＰＣ分析において、自己相関関数を計算すると、音声の特徴パラメータの一つとして、音声のパワーが求められる。

【００５３】音声パワーを示す曲線４０２は、音声波形４０１の音声信号から計算されたパワーである。時間ｔに対するパワーの変化を表わしている。ここで、このパワー情報に対して、しきい値を任意に設定し、入力された音声毎にこのしきい値を超えたかどうかを観測する。この観測は、音声分析部３０３で行う。さらに、複数のしきい値を設け、入力された音声毎にそれぞれのしきい値を超えたかどうかを観測する。例えば、音声パワー曲線４０２の音声の場合は、しきい値ＴＨ１超えているが、しきい値ＴＨ２は超えていない。すなわち、連続的にパワーが、しきい値ＴＨ１とＴＨ２の間にある場合には、感情のレベルを１と見なし、音声分析部３０３は感情レベル１を出力する。

【００５４】つぎに、話者が、感情を込めた強い口調の音声で「あつい」と発声したときの音声波形及び音声パワーをそれぞれ図４（ｃ）及び図４（ｄ）に示す。図４（ｃ）は横軸に時間を、縦軸に音声レベルを示し、図４（ｄ）は横軸に時間を、縦軸に音声パワーを示す。図４（ｃ）に於いて、４０３は音声波形を示す。、図４（ｄ）に於いて、４０４は音声波形４０３の音声信号から計算された音声のパワーであり、時間ｔに対するパワーの変化を表わしている。例えば、音声パワー４０４の場合は、しきい値ＴＨ１超えて、さらに、しきい値ＴＨ２を超えている。すなわち、連続的にパワーが、しきい値ＴＨ２を超えている場合には、感情のレベルを２と見なし、音声分析部３０３は感情レベル２を出力する。この例の場合は、感情レベルを２段階に設定したが、しきい値を増やすことによって、感情レベルをN（Nは整数）段階に設定できる。

【００５５】また、音声分析部３０３からは、感情レベルとともに、音声認識の為の音声の特徴パラメータが時々刻々と音声感情認識部３０４にに入力され、最終的に、音声感情認識部３０４からは音声認識結果「ａｔｕｓｉ」（＝あつい）と感情レベルNを示すデータ３０９

が出力される。

【００５６】次に、図５を用いて本発明に係る音声認識システムのハードウエア構成について説明する。

【００５７】音声を取り込むためのマイク５０１は、カーナビゲーションシステム、携帯型情報端末、ＰＤＡ、ハンドヘルドＰＣ、ゲーム、携帯型翻訳機、並びに、エアコン等の家庭電化製品等では、周囲の雑音を取り込まないために指向性をもたせた指向性マイクである。５０４は、マイク５０１により取り込まれたアナログ音声データを、デジタル音声データに変換するＡ／Ｄ変換器である。

【００５８】音声入力用ボタン５０２は、音声を入力している区間を指定するためのボタンである。ボタンが押されている間、あるいは、ボタンが押された時点から音声が入力されたことをシステムに知らせる。５０５は、音声入力用ボタン５０２と、システムを接続するためのインタフェースである。

【００５９】キー入力用デバイス５０９は、例えば、携帯型情報端末であれば、ペン入力用のデジタイザであり、ハンドヘルドＰＣであれば、キーボードである。また、ファミコンなどのゲーム機であれば、キャラクタ等を操作するキーパッドや、ジョイスティックである。５１０は、キー入力用デバイス５０９と、システムを接続するためのインタフェースである。

【００６０】ＣＰＵ５０３は、カーナビゲーションシステム、携帯型情報端末、ＰＤＡ、ハンドヘルドＰＣ、ゲーム、携帯型翻訳機、並びに、家庭電化製品等のメインシステムの制御と、音声感情認識システムにおける音声認識および感情認識処理を行う。図３に示す本発明の音声感情認識システムの音声分析部３０３、音声感情認識部３０４及びモデル連結部３０８はこのＣＰＵ５０３に設けられる。このＣＰＵ５０３には、ＲＩＳＣマイコンやＤＳＰが用いられるのが、最近の潮流である。

【００６１】ＲＯＭ５０６は、音声認識用単語辞書、音響モデル、発声変形感情モデル、プログラムを格納しておく記憶装置である。また、複数の辞書や、音響モデル、発声変形感情モデルを格納しておくために、メモリカードを用いてもよい。

【００６２】ＲＡＭ５０７は、ＲＯＭ５０６から転送された一部の辞書や、音響モデル、プログラムが格納され、また、音声感情認識処理に必要な必要最小限のワークメモリであり、ＲＯＭ５０６に比べて、通常アクセス時間の短い半導体素子が用いられる。また、ここにはＣＰＵ５０３から音声認識結果及び感情のレベルを示すデータ３０９が入力される。

【００６３】バス５０８は、システムにおけるデータバス、アドレスバス、制御信号バスとして用いられる。

【００６４】音声感情認識結果を出力表示するためのディスプレイ５１２は、ＴＦＴ液晶ディスプレイ等のＬＣＤで構成し、音声認識結果および音声の感情レベルを表示する。５１１は、ディスプレイ５１２と、システムを接続するためのインタフェースである。

【００６５】音声感情認識結果を音で出力するためのスピーカ５１４は、音声認識結果および音声の感情レベルを音声合成して出力する。５１３は、音声認識結果および、音声の感情レベルをテキストから音声合成データに変換処理した後、デジタル音声合成データからアナログ音声信号に変換するＤ／Ａ変換器である。

【００６６】以下、本発明に係る実施形態の一例を、図６及び図８を用いて説明する。

【００６７】本実施形態では、本発明の音声感情認識システムをカーエレクトロニクス製品に適用した場合について説明する。

【００６８】図６（ａ）は本発明による音声感情認識システムをカーエレクトロニクス製品のエアコン操作に利用した場合のブロック図であり、図６（ｂ）はこの音声感情認識システムのおける音声入力例とその認識結果を示す模式図である。

【００６９】図６（ａ）に於いて、６０１は音声入力用マイク、６０２は音声感情認識システム、６０３は音声感情認識結果や、話者との双方向のやり取りを行うために、会話形式の文字情報を出力するためのディスプレイ、６０４は音声感情認識結果や、話者との双方向のやり取りを行うために、会話形式の文字情報を音声合成して出力するためのスピーカである。

【００７０】次に、図６（ｂ）を用いて、話者が発声した音声入力例と、音声感情認識システムが出力した認識結果例を説明する。

【００７１】６０５は話者が音声感情システム６０２に対して、普通の発声で、「暑い」と発声した場合の音声感情システム６０２の認識結果である「暑い＝感情レベル３」を示す。次に、６０６は話者が音声感情システムに対して、強い調子で、「暑い」と発声した時の音声感情システム６０２の認識結果であり、「暑い＝感情レベル５」を示す。

【００７２】また、６０７は話者が音声感情システムに対して、普通の発声で、「暑い」と発声したときの音声感情システム６０２の認識結果である「少し暑いですか」を示す。次に、６０８は話者が音声感情システムに対して、強い調子で、「暑い」と発声した時の音声感情システム６０２の認識結果である「かなり暑いですか」を示す。

【００７３】さらに、認識結果６０８に対して「はい」と声感情システム６０２に対して発声すると、音声感情システム６０２その認識結果として、「車内を２５℃に設定します」を出力する。実際に、車内が２５℃に設定される。

【００７４】また、他の本実施形態では、本発明の音声感情認識システムをファミリーコンピュータ（登録商標）等のゲーム製品に適用した場合について説明する。

【００７５】図８（ａ）から図８（ｄ）は、ファミリコンピュータ等のゲーム機のキャラクタの操作において、音声感情認識システムを用いた一例であり、音声によるインターフェースの音声入力例及び認識結果による動作例を示す模式図である。

【００７６】図８に於いて、８０１、８０２、８０９及び８１０は、ゲーム機本体のディスプレイやゲーム機が接続されたＴＶ等の画面である。

【００７７】図８（ａ）に於いて、ゲームに登場するキャラクター８０５は、例えば、画面８０１に向かって左から右へと進んでいる。この操作を音声感情認識を使って行う。そこで、操作者（話者）は、発声例８０３に示すように「進め」と普通の音声で発声する。音声感情認識システムは、進め（ｓｕｓｕｍｅ）を認識し、さらに、感情レベルを認識する。例えば、このゲーム機における音声感情認識システムでは、感情レベルを５段階に設定したとすると、感情レベル＝３と認識する。そこで、ゲーム機本体側のシステムでは、キャラクター８０５をキャラクター８０６の位置へ移動する。

【００７８】図８（ｂ）に於いて、ゲームに登場するキャラクター８０７は、例えば、画面８０２に向かって左から右へと進んでいる。この操作を音声感情認識を使って行う。そこで、操作者（話者）は、音声例８０４に示すように、「進め」と強い調子で発声する。音声感情認識システムは、進め（ｓｕｓｕｍｅ）を認識し、さらに、感情レベルを認識する。例えば、このゲーム機における音声感情認識システムでは、感情レベルを５段階に設定したとすると、感情レベル＝５と認識する。そこで、ゲーム機本体側のシステムでは、キャラクター８０７はキャラクター８０８の位置へ大きく移動する。ここで、キャラクターの移動量は、認識された音声「進め」の感情レベルに比例する。

【００７９】図８（ｃ）に於いては、ゲームに登場するキャラクター８１３は、例えば、画面８０９に向かって左から右へと進んでいる。このとき、前方に障害物８１６が現われたとする。そこで、この障害物８１６を飛び超えなくてはならない。この操作を音声感情認識を使って行う。そこで、操作者（話者）は、音声例８１１に示すように「ジャンプ」と普通の音声で発声する。音声感情認識システムは、ジャンプ（ｊｙａｎｐｕ）を認識し、さらに、感情レベルを認識する。例えば、このゲーム機における音声感情認識システムでは、感情レベルを５段階に設定したとすると、感情レベル＝３と認識する。そこで、ゲーム機本体側のシステムでは、キャラクター８１３をキャラクター８１４の位置へ移動し、さらに、キャラクター８１５の位置へ移動する。

【００８０】図８（ｄ）にに於いては、ゲームに登場するキャラクター８１７は、例えば、画面８１０に向かって左から右へと進んでいる。このとき、前方に障害物８２０が現われたとする。この障害物８２０は、画面８０９のときの障害物８１６よりも大きい。そこで、この障害物８２０を高く飛び超えなくてはならない。この操作を音声感情認識を使って行う。そこで、操作者（話者）は、音声例８１２に示すように「ジャンプ」と強い調子で発声する。音声感情認識システムは、ジャンプ（ｊｙａｎｐｕ）を認識し、さらに、感情レベルを認識する。例えば、このゲーム機における音声感情認識システムでは、感情レベルを５段階に設定したとすると、感情レベル＝５と認識する。そこで、ゲーム機本体側のシステムでは、キャラクター８１７をキャラクター８１８の位置へ大きく移動し、さらに、キャラクター８１９の位置へ移動する。キャラクターの移動量は、認識された音声「ジャンプ」の感情レベルに比例する。

【００８１】

【発明の効果】本発明によれば、カーナビゲーションシステム、小型情報システム、ゲームに用いられる音声認識システムにおいて、登録した辞書の単語の文字列を音声で認識するとともに、音声認識された単語において、話者の音声が持つ感情のレベルを認識することができる音声感情認識システムを提供することができる。

【００８２】また、本発明によれば、音声認識を用いたカーナビゲーションシステム、小型情報システム、ゲームにおいて、音声の感情レベルを認識できることから、限られた単語数においても、音声認識によるインタフェースのバリエーションを増やすことができ、良好な音声認識インタフェースを実現することができる。

【図面の簡単な説明】

【図１】本発明に係る音声感情認識システムの一実施例を示す示すブロック図である。

【図２】図１に示す音声感情認識システムの音響モデルおよび発声変形感情モデルを説明するための模式図である。

【図３】本発明に係る音声感情認識システムの他の実施例を示すブロック図である。

【図４】図３に示す音声感情認識システムの音声波形および音声パワーと感情レベルの関係を説明するための模式図である。

【図５】本発明のハードウエア構成を示すブロック図である。

【図６】本発明の音声認識感情システムを適用したカーナビゲーションシステムにおける音声によるインタフェースの音声入力例および認識結果例を示す模式図である。

【図７】従来の携帯型翻訳機のブロック図である。

【図８】本発明の音声認識感情システムを適用したゲーム機における音声によるインタフェースの音声入力例および認識結果による動作例を示した模式図である。

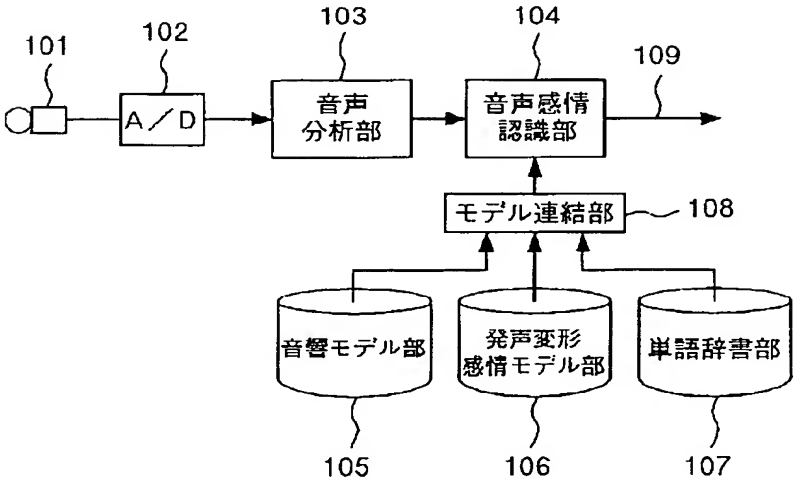
【符号の説明】

１０１、３０１、５０１、６０１…マイク、１０２、３０２、５０４…Ａ／Ｄ変換器、１０３、３０３…音

声分析部、104、304…音声感情認識部、108、308…モデル連結部、105、305…音響モデル部、106…発声変形感情モデル部、107、307…単語辞書部、201…HMM音響モデル連結における「あ」の状態、202…HMM音響モデル連結における「つ」の状態、203…HMM音響モデル連結における「い」の状態、204…状態「あ」から状態「あ」へ遷移する確率、205…状態「つ」から状態「つ」へ遷移する確率、206…状態「い」から状態「い」へ遷移する確率、207…状態「あ」から状態「つ」へ遷移する確率、208…状態「つ」から状態「い」へ遷移する確率、209…状態「い」から他の状態へ遷移する確率、211…状態「あ」の出力確率の連続分布、212…状態「つ」の出力確率の連続分布、213…状態「い」の出力確率の連続分布、214…HMM発声変形感情モデルにおける状態「あ」の出力確率の連続分布、221…HMM発声変形感情モデルにおける状態「あ」の出力確率の連続分布、222…HMM発声変形感情モデルにおける状態「つ」の出力確率の連続分布、223…HMM発声変形感情モデルにおける状態「い」の出力確率の連続分布、502…ボタン、503…CPU、506…ROM、507…RAM、509…キー、602…音声感情認識システム。

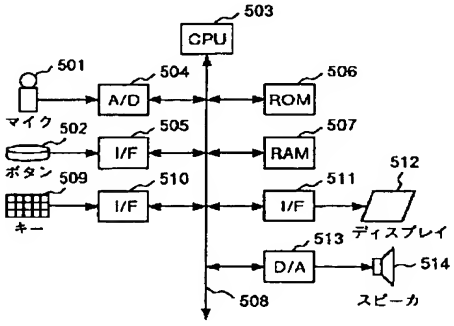
【図1】

図　1



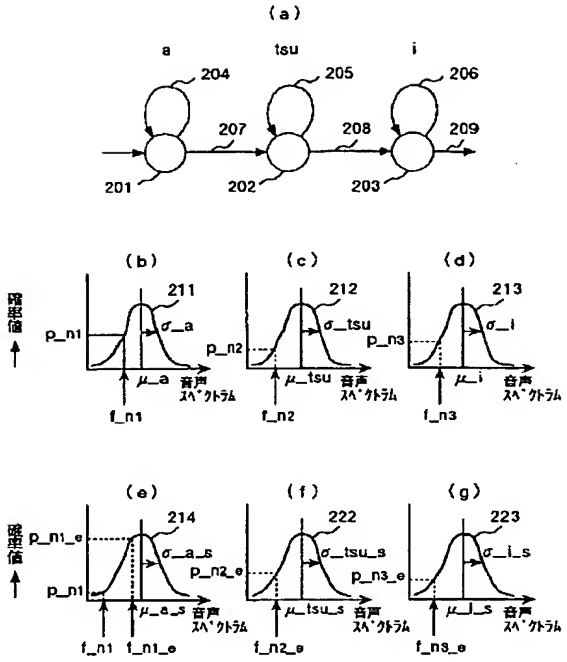
【図5】

図　5



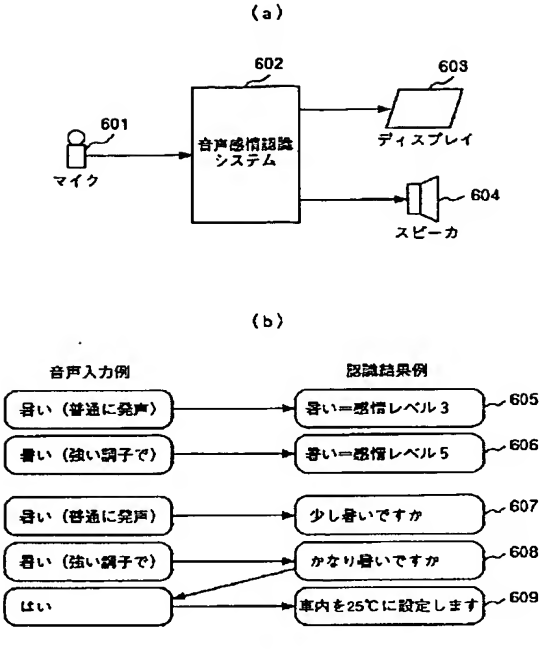
【図２】

図　２



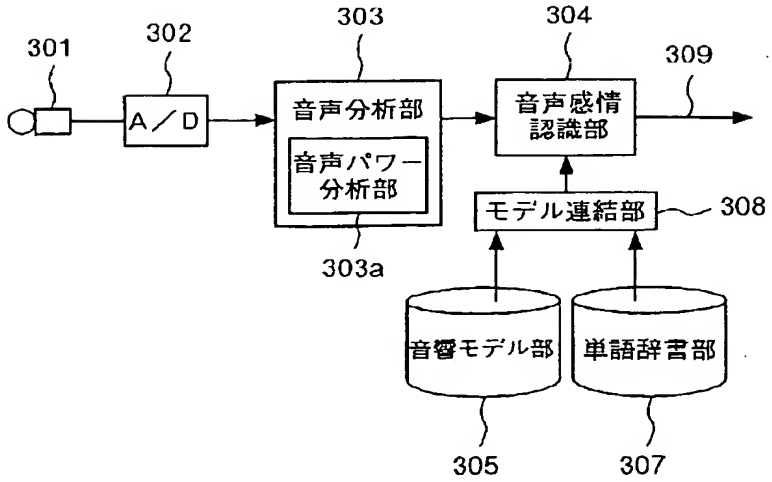
【図６】

図　６



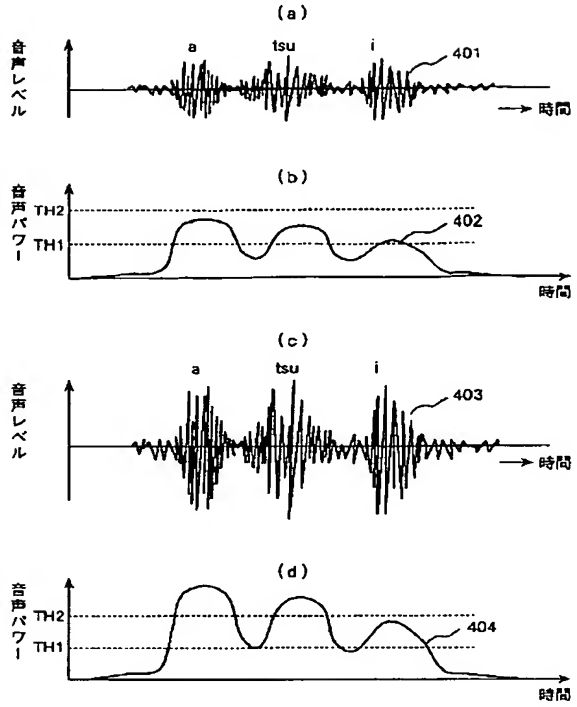
【図３】

図　３



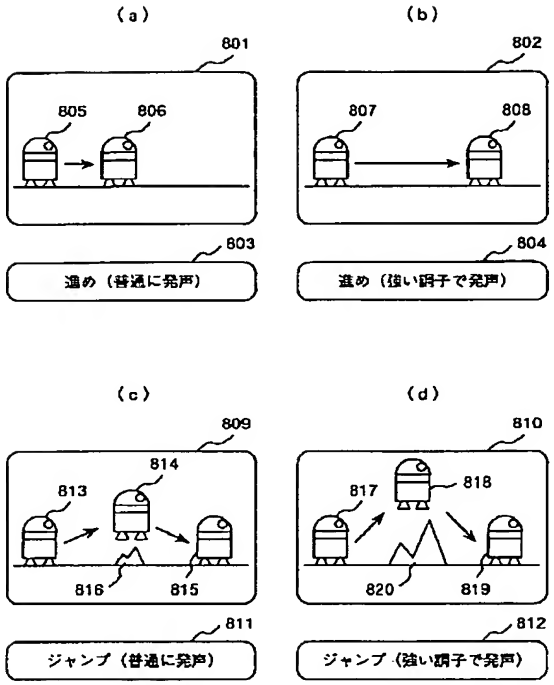
【図4】

図　4



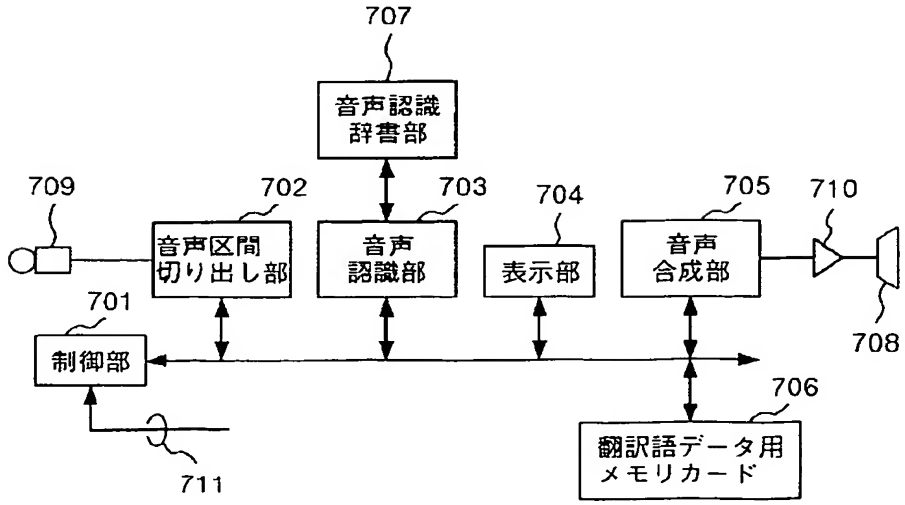
【図8】

図　8



【図7】

図　7



フロントページの続き

(72)発明者　近藤　和夫
東京都小平市上水本町五丁目20番１号株式会社日立製作所半導体事業部内
(72)発明者　大淵　康成
東京都国分寺市東恋ケ窪一丁目280番地株式会社日立製作所中央研究所内
(72)発明者　塔下　哲司
東京都小平市上水本町五丁目20番１号株式会社日立製作所半導体事業部内
(72)発明者　石川　泰代
東京都小平市上水本町五丁目22番１号株式会社日立マイコンシステム内